# Image classification using Markov random fields with two new relaxation methods : deterministic pseudo annealing and modified metropolis dynamics

Zoltan Kato, Josiane Zerubia, Marc Berthod, Jean-Paul Stromboni

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Sophia Antipolis
B.P. 109
06561 Valbonne Cedex
France
Tél.: 93 65 77 77

Rapports de Recherche

N°1606

*Programme 4*
*Robotique, Image et Vision*

# IMAGE CLASSIFICATION USING MARKOV RANDOM FIELDS WITH TWO NEW RELAXATION METHODS:DETERMINISTIC PSEUDO ANNEALING AND MODIFIED METROPOLIS DYNAMICS

Zoltan KATO, Josiane ZERUBIA
Marc BERTHOD

Novembre 1991

# Image classification using Markov Random Fields with two new relaxation methods: Deterministic Pseudo Annealing and Modified Metropolis Dynamics

# Classification d'image à l'aide des champs de Markov avec deux nouvelles méthodes de relaxation: Pseudo-Recuit déterministe et Dynamique de Metropolis modifiée

Zoltan Kato, Josiane Zerubia, Marc Berthod
INRIA, Sophia-Antipolis
2004 Route des Lucioles
06565 Valbonne France
Tel (33) 93 65 78 57
Fax (33) 93 65 76 43

Jean-Paul Stromboni
Université de Nice Sophia-Antipolis
Laboratoire Signaux et Systèmes, URA 1376 CNRS,
41, Boulevard Napoléon III, 06041 Nice France

**Abstract**

In this paper, we present two relaxation techniques: Deterministic Pseudo-Annealing (DPA) and Modified Metropolis Dynamics (MMD) in order to do image classification using a Markov Random Field modelization. For the first algorithm (DPA), the a posteriori probability of a tentative labeling is generalized to continuous labeling. The merit function thus defined has the same maxima under constraints yielding probability vectors. Changing these constraints convexify the merit function. The algorithm solve this unambigous maximization problem and then tracks down the solution while the original constraints are restored yielding a good even if suboptimal solution to the original labeling assignment problem. As for the second method (MMD), it is a modified version of the Metropolis algorithm: at each iteration the new state is chosen randomly but the decision to accept it is purely deterministic. This is of course also a suboptimal technique which gives faster results than stochastic relaxation. These two methods have been implemented on a Connection Machine CM2 and simulation results are shown with a synthetic noisy image and a SPOT image. These results are compared to those obtained with the Metropolis algorithm, the Gibbs sampler and ICM (Iterated Conditional Mode).

**Résumé**

Dans ce rapport, nous présentons deux méthodes de relaxation: un Pseudo-Recuit deterministe (DPA) et une Dynamique de Metropolis modifiée (MMD) appliqués à la classification d'image à l'aide des champs de Markov. Pour le premier algorithme (DPA), la probabilité a posteriori d'un étiquetage d'essai est généralisée à un étiquetage continu. La fonction de mérite ainsi définie présente les mêmes maxima sous des contraintes conduisant aux vecteurs de probabilité. La modification de ces contraintes permet de rendre convexe la fonction de mérite. L'algorithme resoud alors ce probléme non-ambigu de maximisation et recherche la solution alors que les contraintes originelles sont restaurées ce qui conduit à une bonne solution bien que sous-optimale. Pour la deuxième méthode (MMD), c'est une variante de l'algorithme de Metropolis: à chaque itération, le nouvel état est choisi aléatoirement mais la règle d'accepter cet état est déterministe. Bien sûr, c'est une méthode sous-optimale qui est plus rapide que les techniques stochastiques. Les deux algorithmes sont mis en œuvre sur la machine à connexions CM2 et testés sur une image synthétique et sur une image SPOT. Nous les avons comparés à l'algorithme de Metropolis, l'échantillonneur de Gibbs et l'ICM (Iterated Conditional Mode).

**Key Words**

Image Classification, Markov Random Fields, Relaxation algorithms.

**Mots Clefs**

Classification d'image, Champs de Markov, Algorithmes de relaxation.

# Contents

**1 Introduction** 4

**2 Deterministic Pseudo-Annealing** 4

2.1 Probabilistic Modelization . . . . . 5

2.2 Changing the Problem to Approximately Solve it . . . . . 7

**3 Modified Metropolis Dynamics** **9**

3.1 The Model . . . . . 9

3.2 The algorithm . . . . . 10

**4 Simulation results** **11**

4.1 Implementation on a Connection Machine CM2 . . . . . 11

4.2 Comparison with other methods . . . . . 11

4.3 Evaluation of the performances . . . . . 17

**5 Conclusion** **17**

# 1 Introduction

Markov Random Fields (MRF) have become more and more popular during the last few years for many low-level tasks in computer vision. A good reason for that is that such a modelization is the one which requires the less a priori information on the world model. As a matter of fact, the simplest statistical model for an image consists of the probabilities of classes, or grey levels for isolated pixels. The knowledge on the dependencies between nearby pixels is much more powerful, and imposes few constraints. In a way, it is difficult to conceive a more general model, even if it is not easy to determine the values of the parameters which specify a MRF. Another good reason is of course the Hammersley-Clifford theorem, reported for example in [12], which considerably eased, as stressed in [1], the determination of these parameters, by allowing to specify the model either by conditional or joint probabilities.

Many standard computer vision problems, specifically early vision ones such as image classification [8], can thus be expressed quite naturally as combinatorial optimization problems. Direct optimization is not tractable even in the smallest cases. Many heuristics have been proposed to solve them : Iterated Conditional Modes [6, 15], Graduated Non-Convexity [7, 19], Mean Field Annealing [11, 21]. Simulated Annealing [12, 17, 18], Dynamic Programming [9] ....

We propose here two variations on annealing. The first approach, which we propose to call Deterministic Pseudo Annealing (DPA) is indeed related to Relaxation Labeling, a quite popular framework for a variety of computer vision problems [10, 14, 20]. The basic idea is to introduce weighted labelings, which assign a weighted combination of labels to any object, or site, to be labeled, and then to build a merit function of all the weighted labels in such a way as this merit function takes the values of the probability of a global assignment of labels (up to a monotonic transform) for any weighted labeling which assigns the value 1 to one label and 0 to the others at any site. Besides, these values are the only extrema of this function, under suitable constraints. DPA consists of changing the constraints so as to convexify this function, find its unique global maximum, and then track down the solution, by a continuation method, until the original constraints are restored, and a discrete labeling can be obtained.

The second approach (MMD) is a modified version of the Metropolis algorithm [18]: for each iteration, a global state is chosen at random (with a uniform distribution) and for each site, the decision of accepting the new state is deterministic.

These two new methods have been implemented on a Connection Machine CM2 and compared to ICM, Metropolis algorithm and Gibbs sampler on noisy synthetic data as well as SPOT images.

# 2 Deterministic Pseudo-Annealing

Herein, we are interested with the following general problem : we are given a set of units (or sites) $\mathcal{S} = S_i$, $1 \leq i \leq N$, each of which may belong to any one of $M$ classes, or

equivalently take any label from 1 to $M$. We are also given a MRF on these units, defined as usual by a graph, and the so-called clique potentials. We denote the graph by $G$, an edge of $G$ by $E_{ij}$, the set of vertices (or sites) connected to a given vertex $S_i$ by $V_i$. Let $c$ denote a clique of $G$, and $\mathcal{C}$ the set of all cliques of $G$. Also $C_i = \{c : S_i \in c\}$. The number of sites in the clique is its degree : $\deg(c)$, and $\deg(G) = \max_{c \in \mathcal{C}} \deg(c)$.

A global discrete labeling $L$ assigns one label $L_i$ ($1 \leq L_i \leq$M) to each site $S_i$ in $\mathcal{S}$. The restriction of $L$ to the sites of a given clique $c$ is denoted by $L_c$. The definition of the MRF is completed by the knowledge of the clique potentials $V_{cL}$ (shorthand for $V_{cL_c}$) for every $c$ in $\mathcal{C}$ and every $L$ in $\mathcal{L}$, where $\mathcal{L}$ is the set of the $M^N$ discrete labelings (recall that $M$ is the number of possible labels, which is assumed to be the same for any site for simplicity, and $N$ is the number of sites).

## 2.1 Probabilistic Modelization

The nice result of Hammersley-Clifford is that the probability of a given labeling $L$ may be computed quite easily (assuming deg (C) is small, at most 2 or 3), by :

$$P(L) = \frac{\prod_{c \in \mathcal{C}} \exp(-V_{cL})}{Z} \tag{1}$$

where $Z$, the partition function, is a normalizing factor such that :

$$\sum_{L \in \mathcal{L}} P(L) = 1$$

We assume here that the sufficient positivity condition $P(L) > 0$ is met.

The basic problem, for most applications, is to find the labeling $L$ which maximizes $P(L)$, knowing that exhaustive search of all the labelings is strictly intractable.

Before explaining the method we propose, its algorithmic realization and the type of modelization it can be derived from, it is necessary to give more details about the way Bayesian Modeling behaves with Markov Random Fields.

First of all, it is important to notice that for most applications, the information available stems from two different sources : a priori knowledge about the restrictions that are imposed on the simultaneous labeling of connected neighbor units, and observations that have been made, for a given occurence of the problem, on these units.

The first source is generic, and typically referred to as the World Model. For example, Discrete Relaxation relies on the knowledge of allowed couples, or n-tuples of labels between neighbouring sites. This type of knowledge may be more detailed, and reflect statistical dependencies between the labels of neighbouring sites, thus defining a Markov Random Field. For example, when dealing with images, the knowledge of the likelihood of configurations of nearby pixels may take the form of a MRF with cliques of order 1 to 2 (4-connectivity) or order 1 to 4 (8-connectivity). The other source of information consists

of the observations. Combining these two sources of information may be achieved in different ways; Bayiesian Modeling, whether strictly applied or not, comes in very naturally at this stage. Let us assume, for simplicity, that the observations consist of the grey levels (or any other scalar or vector quantities) of the pixels in an image : $y_i$ is thus the grey level for pixel (or site, or unit) $S_i$. Let $Y = (y_1 \dots y_N)^t$. $Y$ represents here the observed image. A very general problem is to find, given for example a first order MRF on these pixels (i.e. knowing the statistics of couples of labels with 4-connectivity), the labeling $L$ which maximizes $P(L/Y)$. Bayes theorem tells us that $P(L/Y) = P(Y/L)P(L)/P(Y)$. Actually P(Y) does not depend on the labeling L, and plays exactly the same role as $Z$ in Equation (1), at least as far as testing the likelihood of the MRF is not at hand. We have now to assume that we are able to model the noise process. Standard assumptions, which roughly amount to white invariant noise, are that :

$$P(Y/L) = \prod_{i=1}^{N} P(y_i/L) = \prod_{i=1}^{N} P(y_i/L_i) \tag{2}$$

As for P(L), it is taken care of by the MRF modeling the a priori world model, as in Equation (1). It is then easy to see that the a posteriori probability, which we are trying to maximize, is given by :

$$P(L/Y) \simeq \prod_{i=1}^{N} P(y_i/L_i) \prod_{c \in \mathcal{C}} \exp(-V_{cL}) \tag{3}$$

(from now on, we will use $\simeq$ instead of $=$ whenever we drop a scaling factor). It is obvious from this expression that the a posteriori probability also derives from a Markov Random Field, with cliques of order 1 and 2 (and not only 2 as for the a priori probability). The energies of cliques of order 1 directly reflect the probabilistic modeling of labels without context, which would be used for classifying, or labeling the pixels independently. This equivalence was used in [4, 5] for initializing a continuous labeling before relaxation. But it is not very difficult to prove that it is always possible, by suitable shifts on the clique potentials to keep only the potentials of maximal cliques. The procedure to do so directly derives from the proof of the Hammersley-Clifford theorem given by [6]. The problem at hand is thus strictly equivalent to maximizing :

$$f(L) = \sum_{c \in \mathcal{C}} W_{cL} \tag{4}$$

where $W_{cL} = -V_{cL}$ and let $\hat{L}$ be the corresponding labeling. As we shall see, this transformation may not be really necessary, but it will simplify some results. The following property is more interesting : shifting every clique potential of a given clique by the same quantity is equivalent to scaling all the $P(L)$'s by a given factor, or equivalently to change the normalization factor $Z$ of Equation (1). Thus, the maximization problem is not changed. For example, it will be possible to shift all the $W_i$'s so that they all become non-negative, or even positive, without changing the solution to the problem.

## 2.2 Changing the Problem to Approximately Solve it

Several approaches have been proposed to find at least a reasonably good labeling. One of the best known is probably simulated annealing [12]. But other more "algorithmic" approaches are worth mentioning : Iterated Conditional Modes (ICM) [6], or dynamic programming [9] for example. Even former works on relaxation labelings, already mentioned, were (knowingly or not) a way to tackle this problem. We start from one such work [5]. The key point here is to cast this discrete, combinatorial, optimization problem into a more comfortable maximization problem on a compact subset of $\mathcal{R}^N$. Let us define a real function $f(X)$ ($X \in \mathcal{R}^{\mathcal{NM}}$) as follows :

$$f(X) = \sum_{c \in \mathcal{C}} \sum_{l_c \in L_c} W_{cl_c} \prod_{j=1}^{\deg(c)} x_{c_j, l_{c_j}} \tag{5}$$

where $c_j$ denotes the $j^{th}$ site of clique c, and $l_{c_j}$ the label assigned to this site by $l_c$. It is clear from Equation (5) that $f$ is a polynomial in the $x_{i,k}$'s; the maximum degree of $f$ is the maximum degree of the cliques. If we assume for simplicity that all the cliques have the same degree $d$ (this is easily true with 4-neighbourhoods on images, after suitable shifts on the coefficients), then $f$ is a homogeneous polynomial of degree $d$. This is by no means necessary in what follows but will alleviate the notations.

Moreover, it is clear that $f$ is linear with any $x_{i,k}$. Let us now restrict X to $\mathcal{P}_{NM}$, a specific compact subset of $\mathcal{R}_{NM}$ defined by the following constraints :

$$\forall i,k \ : x_{i,k} \geq 0 \tag{6}$$

$$\forall i \ : \sum_{k=1}^{M} x_{i,k} = 1 \tag{7}$$

It turns out that, generically, the maxima of $f$ on $\mathcal{P}_{NM}$ (a convex polytope) are all located on the border, i.e. if $X^\star$ is such a maximum, then

$$\forall i, \exists k \ : x^\star_{ik} = 1, \ l \neq k \Rightarrow x^\star_{il} = 0 \tag{8}$$

The reason for that is that f is a concave function (actually a parabola) of $x_{ik}, \forall i,k$. Thus, if $\frac{\partial f}{\partial ik} = 0$, then f is minimal; the only way for f to be maximal is that the point is on the border ($x_{ik} = 0 \ or 1$), and the gradient points outside.

Actually, it may happen that for some $i$, two or more labels may be indifferent, but the key point is that there is no maximum *inside*, but degenerate ones.

Thus, any maximum of $f$ on $\mathcal{P}_{NM}$ directly yields a discrete labeling, and the absolute maximum of $f$ yields the solution to our problem. The difficulty is of course that $f$ is not concave but convex with a tremendous number of such maxima, and that any standard gradient technique will usually lead to a local maximum, and not to the absolute maximum. It is thus vital to find a good starting point before applying such a technique.

The basic idea in DPA is to change temporarily the subset on which $f$ is maximized so that $f$ becomes concave, to maximize $f$, and to track this maximum while slowly changing the constraints until the original ones are restored so that a discrete labeling can be deduced.

First of all, when $c = 2$ (cliques of order 2), $f$ is a quadratic form, and can always be written as $f = X^t AX$, where $A$ is an $NM \star NM$ symmetric matrix. Besides, after suitable shift, $A$ has non-negative entries. After Perron-Frobenius [3], $A$ has a unique real non-negative eigenvector, which is strictly positive, the corresponding eigenvalue being positive and equal to the spectral radius; besides any other eigenvalue has a smaller modulus. Actually, this is a generic case, which admits degeneracies, but they can be dealt with easily and are not considered here (see [4] for more details). This eigenvector maximizes $f$ under constraints different from the preceding ones :

$$\forall i,k \ : x_{i,k} \geq 0 \tag{9}$$

$$\forall i \ : \sum_{k=1}^{M} x_{i,k}^2 = 1 \tag{10}$$

We call $Q^{NM,d}$ the compact subset of $\mathcal{R}^{NM}$ so defined.

Moreover, the vector $X$ can be obtained very efficiently by using (for example) the iterative power method : start from any $X^0$, and apply

$$X^{n+1} = \frac{AX^n}{\|AX^n\|_{L^2}} \simeq AX^n \tag{11}$$

A fundamental point is the following : if $f$ is a polynomial with non-negative coefficients and maximum degree $d$, then $f$ has a unique maximum on $Q^{NM,d}$ (with again possible degeneracies as mentioned when $d = 2$). See [2] for a complete proof.

The iterative power method can be readily extended, becoming : select $X = X^0$, and apply

$$X^{n+1} \simeq (\bigtriangledown f(X^n))^{\frac{1}{d-1}} \tag{12}$$

This simply means that, at each iteration, we select on the pseudo-sphere of degree $d$ the point where the normal is parallel to the gradient of $f$. Obviously, the only stable point is singular, and thus is the maximum we are looking for. We have only proved experimentally that the algorithm does converge very fast to this maximum.

This procedure, already suggested in [5] yields a maximum which, as in the case $d = 2$, is inside $Q^{NM,d}$ (degeneracies apart), and thus does *not* yield a discrete labeling. The second key point is now to decrease $d$ down to 1. More precisely, we define an iterative procedure, as follows :

- set $\beta := d$, select some $X$

- *while* $(\beta > 1)$ *do*
  find $X^\star$ which maximizes $f$ on $Q^{NM,\beta}$, starting from $X$
  decrease $\beta$ by some quantity
  project $X^\star$ on the new $Q^{NM,\beta}$ giving $X$

- *od*

- for each $S_i$, select the label with value 1

This iterative decrease of $\beta$ can be compared, up to a point to a cooling schedule, or better to a Graduated Non-Convexity strategy [7].

The last step (projection) is necessary, as changing $\beta$ changes $Q^{NM,\beta}$. Actually, the normalization performed at the first iteration of the maximization process, for any $\beta$, takes care of that. On the other hand, the process defined by Equation (11) cannot be applied when $\beta = 1$. Maximization in that case has been thoroughly studied by [10]. But it is simpler to stop for some $\beta$ slightly larger than 1 (eg. 1.1), as experiments confirm that the vector $X^\star$ almost satisfies the constraints in (8), and thus a discrete labeling can easily be deduced.

It is also important to notice that, though shifting the coefficients does not change the discrete problem nor the maximization problem on $\mathcal{P}^{NM}$, it changes it on $Q^{NM,d}$, and thus there is no guarantee that the same solution is reached. Nor is it guaranteed that the procedure converges toward the global optimum; actually, it is not difficult to build simple counterexamples on toy problems. Experiments show nevertheless that, on real problems, a very good solution is reached and that the speed whith wich $\beta$ is decreased is not crucial : typically, 5 to 10 steps are enough to go from 2 to 1.

# 3 Modified Metropolis Dynamics

## 3.1 The Model

Using the same notations as in Section 2., we want to maximize $P(L/Y)$ as defined in Equation (3). We suppose that $P(y_i/L_i)$ is Gaussian:

$$P(y_i/L_i) = \frac{1}{\sqrt{2\pi}\sigma_{L_i}} \exp\left(-\frac{(y_i - \mu_{L_i})^2}{2\sigma_{L_i}^2}\right) \tag{13}$$

where $\mu_l$ $(1 \leq l \leq M)$ is the mean and $\sigma_l$ $(1 \leq l \leq M)$ is the deviation of the class $l$. The second term is Markovian:

$$P(L) = \exp\left(-\frac{U(L)}{T}\right) \;=\; \exp\left(-\frac{1}{T}\sum_{c\in\mathcal{C}} V_{cL}\right) \tag{14}$$

Choosing $T = 1$, we get:

$$\begin{aligned}
\hat{L} &= \max_{L \in \mathcal{L}} (\log P(Y/L) + \log P(L)) \\
&= \max_{L \in \mathcal{L}} \left( \sum_{i=i}^{N} - \left( \log \sqrt{2\pi}\sigma_{L_i} + \frac{(y_i - \mu_{L_i})^2}{2\sigma_{L_i}^2} \right) - \sum_{c \in \mathcal{C}} V_{cL} \right) \\
&= \min_{L \in \mathcal{L}} \left( \sum_{i=i}^{N} \left( \log \sqrt{2\pi}\sigma_{L_i} + \frac{(y_i - \mu_{L_i})^2}{2\sigma_{L_i}^2} \right) + \sum_{c \in \mathcal{C}} V_{cL} \right) \qquad (15)
\end{aligned}$$

Using the above equation, it is easy to define the global and local energies as follows:

$$\mathcal{E}_{glob}(L) = \sum_{i=i}^{N} \left( \log \sqrt{2\pi}\sigma_{L_i} + \frac{(y_i - \mu_{L_i})^2}{2\sigma_{L_i}^2} \right) + \sum_{c \in \mathcal{C}} V_{cL} \qquad (16)$$

$\mathcal{E}_i(L)$ stands for the local energy of the labeling $L$ at the site $S_i$:

$$\mathcal{E}_i(L) = \log \sqrt{2\pi}\sigma_{L_i} + \frac{(y_i - \mu_{L_i})^2}{2\sigma_{L_i}^2} + \sum_{c_i} V_{c_i L} \qquad (17)$$

The estimation of $\hat{L}$ is done through the energy minimization using a relaxation method with a Modified Metropolis Dynamics (MMD)[16].

## 3.2 The algorithm

The proposed algorithm is a modified version of the Metropolis dynamics [18]:

- the choice of the new label state is done randomly using a uniform distribution
- the rule to accept a new state is deterministic.

The parallel algorithm is the following:

1. Pick up randomly an initial configuration $L^0$, with $k = 1$ and $T = T(1)$.
2. Using a uniform distribution, pick up a global state $L^t$ so that
$$\forall i (1 \leq i \leq N) \ : 1 \leq L_i^t \leq M \text{ and } L_i^t \neq L_i^k \ .$$
3. In order to get the convergence of the algorithm, we can partition the entire image into disjoint regions $\mathcal{R}_n$ such that pixels which belong to the same region are conditionally independent given the data of all the other regions:
$$\mathcal{S} = \bigcup_n \mathcal{R}_n \quad \text{and} \quad \mathcal{R}_n \cap \mathcal{R}_m = \emptyset \ \ (n \neq m) \qquad (18)$$

4. For each site $S_i$, the local energy $\mathcal{E}_i(L^{t'})$ is computed using Equation (17) with:

$$L^{t'} = [L_1^k, L_2^k, \ldots, L_{i-1}^k, L_i^t, L_{i+1}^k, \ldots, L_N^k] \tag{19}$$

Calling $\Delta\mathcal{E}_i = \mathcal{E}_i(L^{t'}) - \mathcal{E}_i(L^k)$, a new label state at site $S_i$ is accepted according to the following rule:

$$L_i^{k+1} = \begin{cases} L_i^t & \text{if } \Delta\mathcal{E}_i \leq 0 \text{ or } \Delta\mathcal{E}_i > 0 \text{ and } \alpha \leq \exp\left(-\frac{\Delta\mathcal{E}_i}{T}\right) \\ L_i^k & \text{otherwise} \end{cases}$$

where $\alpha$ is a constant threshold ($\alpha \in (0,1)$), chosen at the beginning of the algorithm.

5. Decrease of the temperature $T = T(k+1)$ ( $k$ = number of iterations) and goto 2 if the number of modified sites $> threshold$.

There is no explicit formula to get the threshold $\alpha$. For image classification, $\alpha$ is chosen nearly equal to zero if the image is very noisy, otherwise $\alpha$ is equal to 0.5 (for more details about the relationship between $\alpha$ and $\Delta\mathcal{E}_{min}$ see [16]).

# 4 Simulation results

## 4.1 Implementation on a Connection Machine CM2

In this section, we briefly describe the architecture of the Connection Machine, a more detailed description can be found in [13]. The Connection Machine is a single instruction multiple data (SIMD) parallel computer with 8K to 64K processors. Each processor is a 1-bit serial processor, with 32K bytes of local memory and a 7MHz clock. For a given application, the user can dynamically define a particular geometry for the set of physical processors that has been attached.

The processor resource can be virtualized (VP Ratio) when the number of data elements to be processed is greater than the number of physical processors. In such a case, several data elements are processed on a single physical processor. Such a data parallelism model architecture is well suited for computer vision.

For the relaxation algorithms described in this paper, we use the data parallelism (one pixel per virtual processor) and the fast local communications (NEWS).

## 4.2 Comparison with other methods

The goal of this simulation is to evaluate the performances of the two algorithms (DPA and MMD) proposed by the authors. As these methods are suboptimal, it is interesting to compare them with another deterministic relaxation scheme ICM [6, 15], known to be

| | VPR | Nb. of Iter | Total time | Time per It. |
|---|---|---|---|---|
| ICM | 2 | 9 | 0.249 sec. | 0.027 sec. |
| Metropolis | 2 | 90 | 26.32 sec. | 0.293 sec. |
| Gibbs | 2 | 86 | 48.51 sec. | 0.564 sec. |
| MMD | 2 | 79 | 2.37 sec. | 0.030 sec. |
| DPA | 2 | 45 | 0.51 sec. | 0.011 sec. |

Table 1: Results on the checkerboard image with 2 classes

| | win | alert | error |
|---|---|---|---|
| ICM | 0.8439 | 0.1561 | 0.1639 |
| Metropolis | 0.9902 | 0.0098 | 0.0156 |
| Gibbs | 0.9866 | 0.0134 | 0.0194 |
| MMD | 0.9989 | 0.0011 | 0.0083 |
| DPA | 0.9843 | 0.0157 | 0.0197 |

Table 2: Objective performance comparison on the checkerboard image

| | VPR | Nb. of Iter | Total time | Time per It. |
|---|---|---|---|---|
| ICM | 8 | 6 | 0.875 sec. | 0.146 sec. |
| Metropolis | 8 | 52 | 61.39 sec. | 1.181 sec. |
| Gibbs | 8 | 57 | 235.68 sec. | 4.135 sec. |
| MMD | 8 | 30 | 5.75 sec. | 0.192 sec. |
| DPA | 8 | 25 | 2.14 sec. | 0.081 sec. |

Table 3: Results on the SPOT image with 4 classes

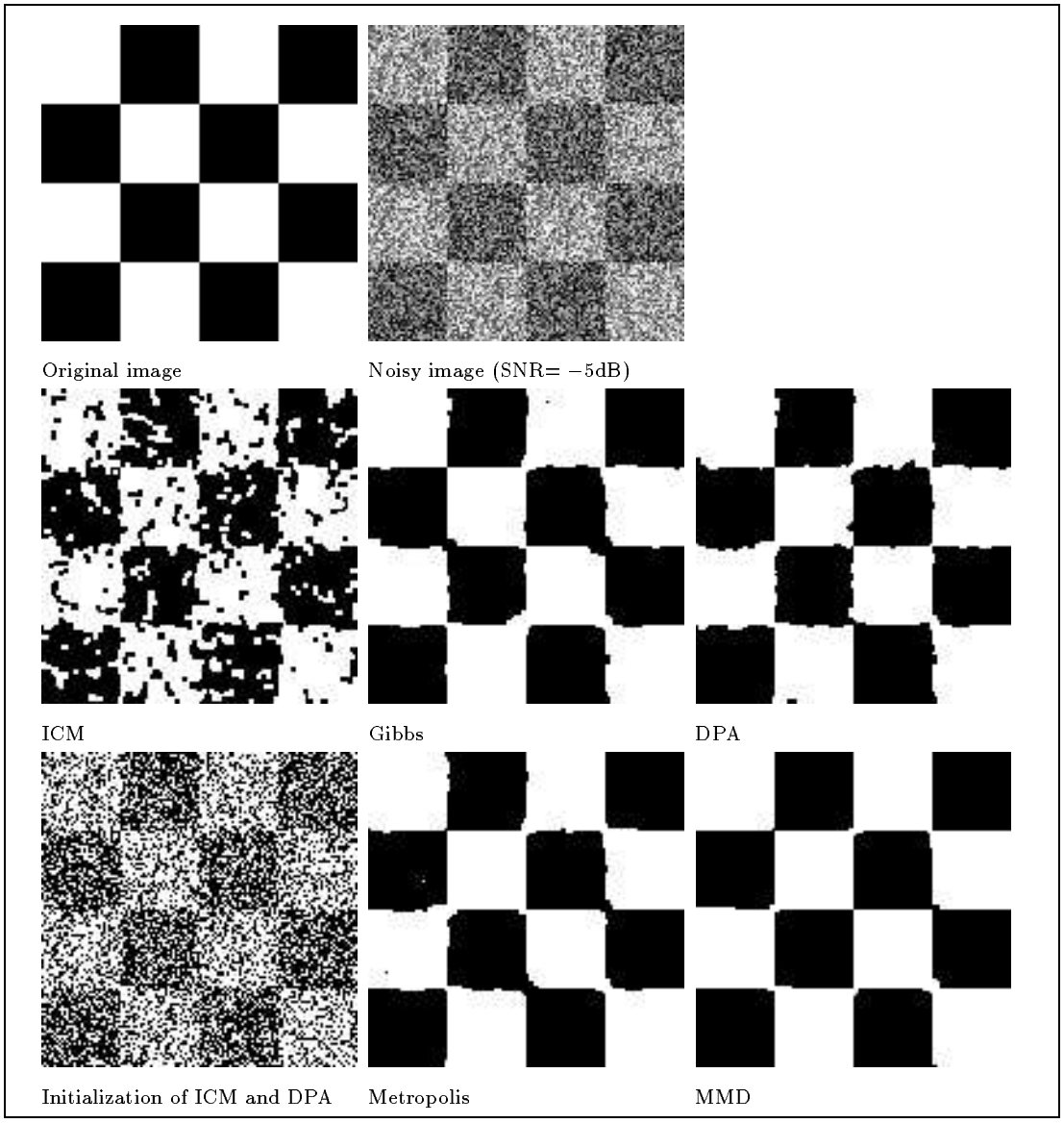


Figure 1: Results with the checkerboard image with 2 classes

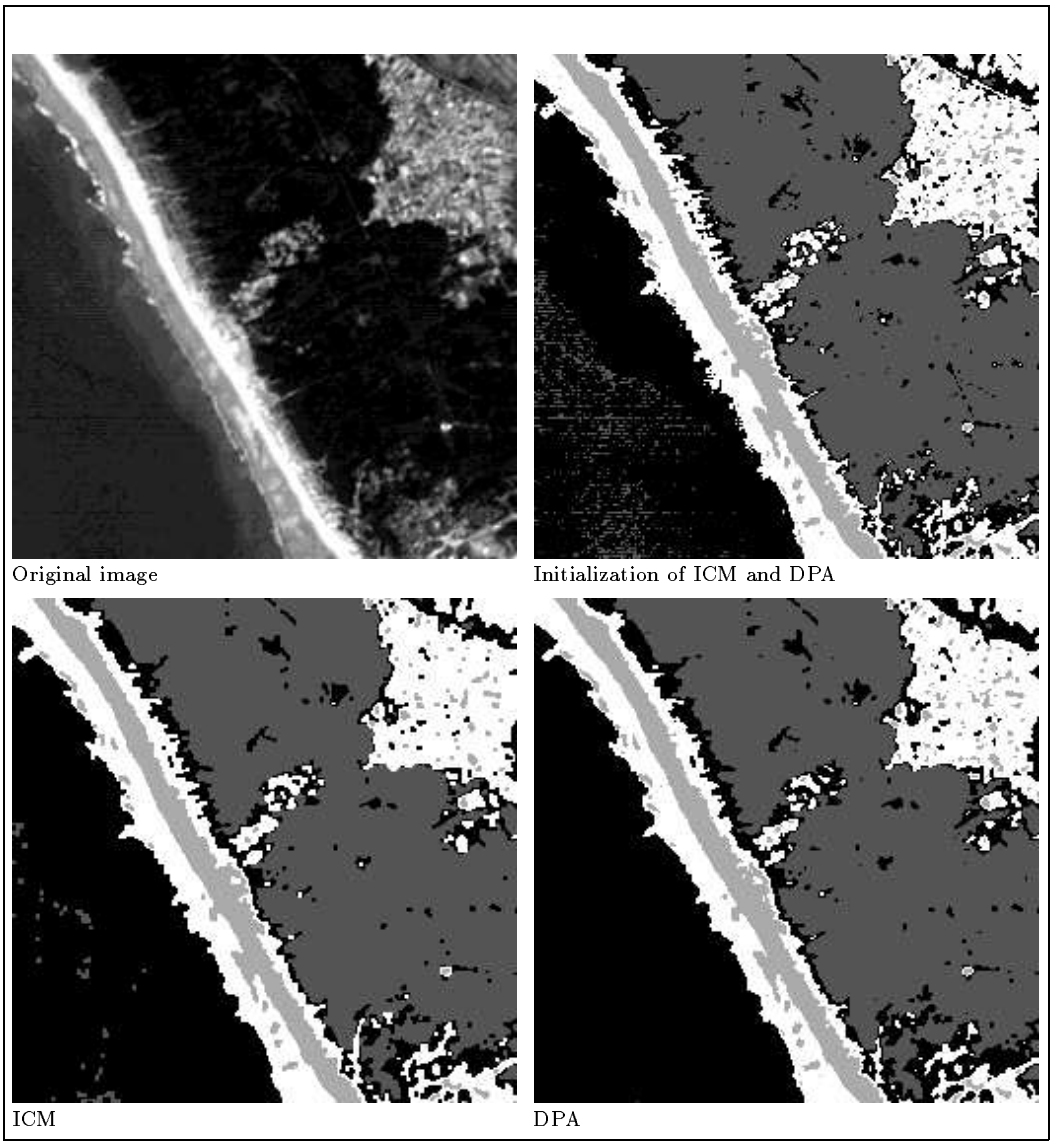


Figure 2: Results with the SPOT image (4 classes)

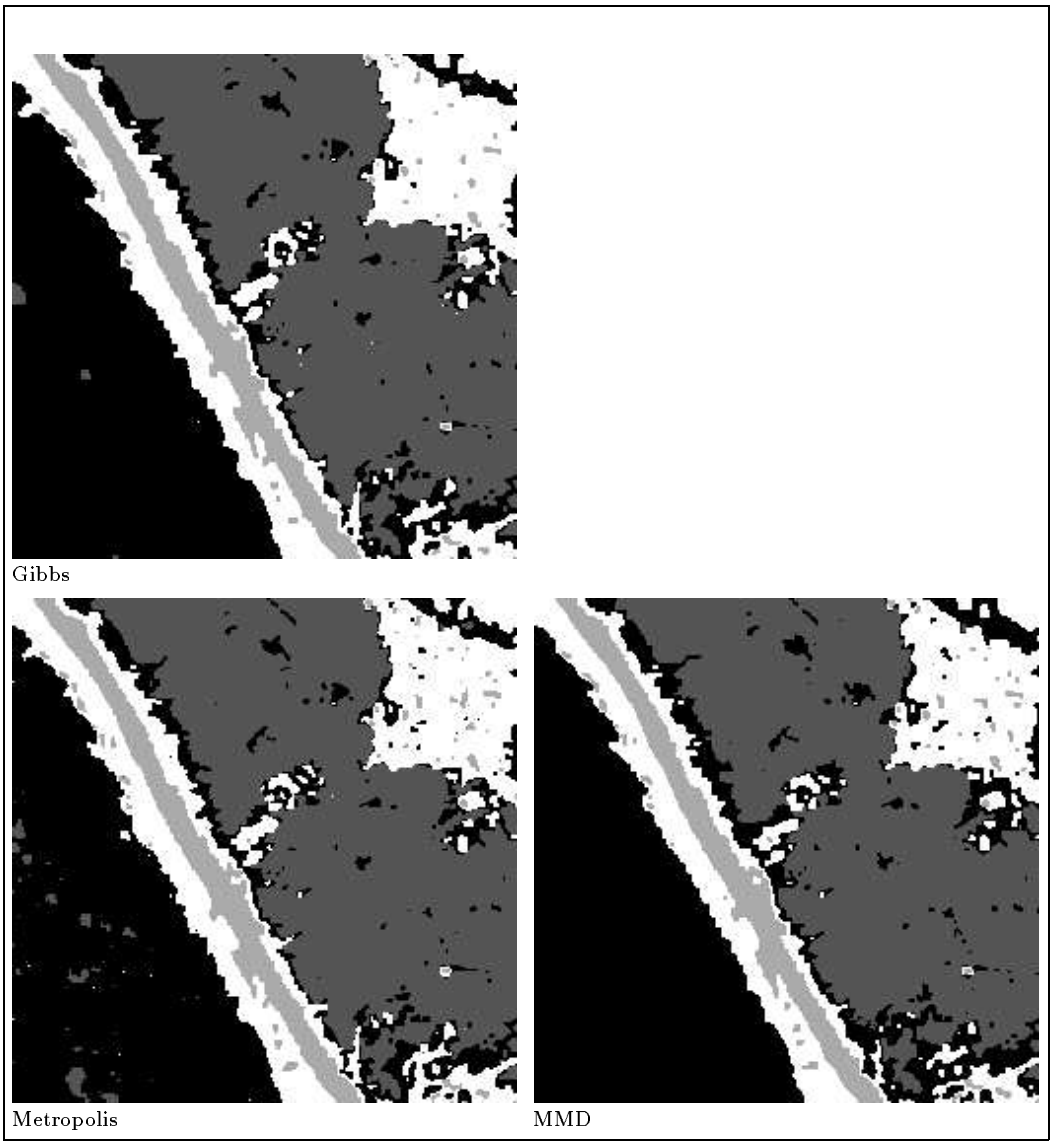


Figure 3: Results with the SPOT image (4 classes)

very fast. Furthermore, it is also worth to compare them with fully stochastic relaxation techniques such as Metropolis dynamics [18] and Gibbs sampler [12].

The tests have been conducted on synthetic as well as real data, in particular SPOT images. Herein, we present the results obtained on a noisy checkerboard image ($128 \times 128$) with a SNR equal to $-5$dB with 2 classes and on a piece of a SPOT image ($256 \times 256$) with 4 classes. The performances are established in two ways:

- objective performances based on the correct classification and computer-time (see tables 1, 2, 3).
- subjective performances based on the visual aspect (see figures 1, 2 3).

Three indicators have been defined for the classification in two classes (see table 2):

$$win = \frac{Number\,of\,white\,pixels\,classed\,white}{Number\,of\,input\,white\,pixels}$$
$$alert = \frac{Number\,of\,white\,pixels\,classed\,black}{Number\,of\,input\,white\,pixels}$$
$$error = \frac{Number\,of\,initially\,black\,pixels\,classed\,white}{Number\,of\,initially\,black\,pixels}$$

For the simulations, the following parameters have been used:

- *Temperature*:
  - *Initial temperature*: $T_0 = 10$ for the synthetic image, $T_0 = 4$ for the SPOT image.
  - $T_{k+1} = 0.95 T_k$ for the decreasing low except for ICM ($T = 1$) and DPA.
- *Mean and deviation of each class*: They are computed using a supervised learning method.

| Image | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ | $\mu_3$ | $\sigma_3$ | $\mu_4$ | $\sigma_4$ |
|---|---|---|---|---|---|---|---|---|
| checkerboard | 0.377 | 0.216 | 0.621 | 0.216 | — | — | — | — |
| SPOT | 0.167 | 0.0244 | 0.120 | 0.0118 | 0.375 | 0.0469 | 0.228 | 0.0489 |

- *Choice of the clique-potentials*:
  - for the checkerboard image (ICM, Gibbs, Metropolis, MMD):

$$V_{\{i,j\}L} = \begin{cases} 0.2 & \text{if } L_i = L_j \\ 2.0 & \text{if } L_i \neq L_j \end{cases}$$

and for DPA:

$$W_{\{i,j\}L} = -V_{\{i,j\}L} = \begin{cases} -0.2 & \text{if } L_i = L_j \\ -2.0 & \text{if } L_i \neq L_j \end{cases},$$

then a shift is used in order to get only positive values (see Equation (4) for more details).

– for the SPOT image (ICM, Gibbs, Metropolis, MMD):

$$V_{\{i,j\}L} = \begin{cases} -0.5 & \text{if } L_i = L_j \\ 0.5 & \text{if } L_i \neq L_j \end{cases}$$

and for DPA:

$$W_{\{i,j\}L} = -V_{\{i,j\}L} = \begin{cases} 0.5 & \text{if } L_i = L_j \\ -0.5 & \text{if } L_i \neq L_j \end{cases},$$

for the same reason.

- *Initialization of the labels*: Random values are assigned to the labels for the initialization of Gibbs, Metropolis and MMD. As for DPA and ICM techniques, the initial labels are obtained using only the Gaussian term in Equation (15) (see figures 1, 2).
- *Initialization of $\alpha$ for MMD*: $\alpha = 0.1$ for the checkerboard image and $\alpha = 0.5$ for the SPOT image.

### 4.3 Evaluation of the performances

First of all, the DPA, MMD and Gibbs sampler give the best results. One should mentioned that, as it is well known, ICM is very sensitive to the initial conditions and may be better results could have been obtained with another initialization. Nevertheless the DPA and ICM algorithms have been initialized with the same data for the simulation.

DPA is the fastest algorithm, on the other hand MMD seems to give slightly better results on these images. Both algorithms offer a good trade-off between quality and computer-time.

Of course, more tests have to be conducted in order to better investigate the possible applications of both algorithms presented this paper.

## 5 Conclusion

In this paper, we have presented two suboptimal methods: Deterministic Pseudo Annealing and Modified Metropolis Dynamics applied to image classification. These techniques (especially the DPA) compares favorably with the classical relaxation algorithms (Metropolis, Gibbs sampler) giving good results and being faster. Furthermore, they are far less dependent on the initial conditions than ICM. This represent a good cost/performance trade-off for early vision problems.

## References

[1] R. Azencott. Markov fields and image analysis. *Proc. AFCET, Antibes*, 1987.

[2] L. Baratchart and M. Berthod. A non-linear generalization of the perron frobenius theorem: an application to the constrained optimization of positive polynomials. *INRIA Research Report (to appear)*, 1991.

[3] A. Berman and R. Plemmons. *Non-negative Matrices in the Mathematical Sciences.* New York, 1979.

[4] M. Berthod. *L'amélioration d'étiquetage : une approche pour l'utilisation du contexte en reconnaissance des formes.* Thèse d'état, Université de Paris VI, December 1980.

[5] M. Berthod. Definition of a consistent labeling as a global extremum. In *Proc. ICPR6*, pages 339–341, Munich, 1982.

[6] J. Besag. On the statistical analysis of dirty pictures. *Jl. Roy. Statis. Soc. B.*, 1986.

[7] A. Blake and A. Zisserman. Visual reconstruction. *MIT Press, Cambridge - MA*, 1987.

[8] P. Chou and C. Brown. The theory and practice of bayesian image labeling. *International Journal of Computer Vision*, 4:185–210, 1990.

[9] H. Derin, H. Elliott, R. Cristi, and D. Geman. Bayes smoothing algorithms for segmentation of binary images modeled by markov random fields. *IEEE trans. on Pattern analysis and mach. intel., Vol 6*, 1984.

[10] O. Faugeras and M. Berthod. Improving Consistency and Reducing Ambiguity in Stochastic Labeling: an Optimization Approach. *IEEE Transactions on PAMI*, 4:412–423, 1981.

[11] D. Geiger and F. Girosi. Parallel and deterministic algorithms for MRFs : surface reconstruction and integration. In *Proc. ECCV90*, Antibes, France, 1990.

[12] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[13] W. D. Hillis. *The Connection Machine.* Cambridge, MA, 1985.

[14] R. Hummel and S. Zucker. On the foundations of relaxation labeling processes. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 5(3), 1983.

[15] F. C. Jeng and J. M. Woods. Compound Gauss - Markov Random Fields for Image Estimation. *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-39:638–697, 1991.

[16] Z. Kato. Modélisation Markovienne en classification d'image. mise en oeuvre d'algorithmes de relaxation. *Rapport de stage de DEA, Université de Nice*, 1991.

[17] P. V. Laarhoven and E. Aarts. Simulated annealing : Theory and applications. *Reidel Pub., Dordrecht, Holland*, 1987.

[18] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. of Chem. Physics, Vol. 21, pp 1087-1092*, 1953.

[19] A. Rangaraja and R. Chellappa. Generalised Graduated Non-Convexity algorithm for Maximum A Posteriori image estimation. *Proc. ICPR*, pages 127–133, Jun. 1990.

[20] R. H. A. Rosenfeld and S. W. Zucker. Scene labeling by relaxation operation. *IEEE Trans. on Systems Mans and Cybernetics*, 1982.

[21] J. Zerubia and R. Chellappa. Mean field approximation using compound Gauss-Markov Random field for edge detection and image restoration. *Proc. ICASSP, Albuquerque, USA*, 1990.